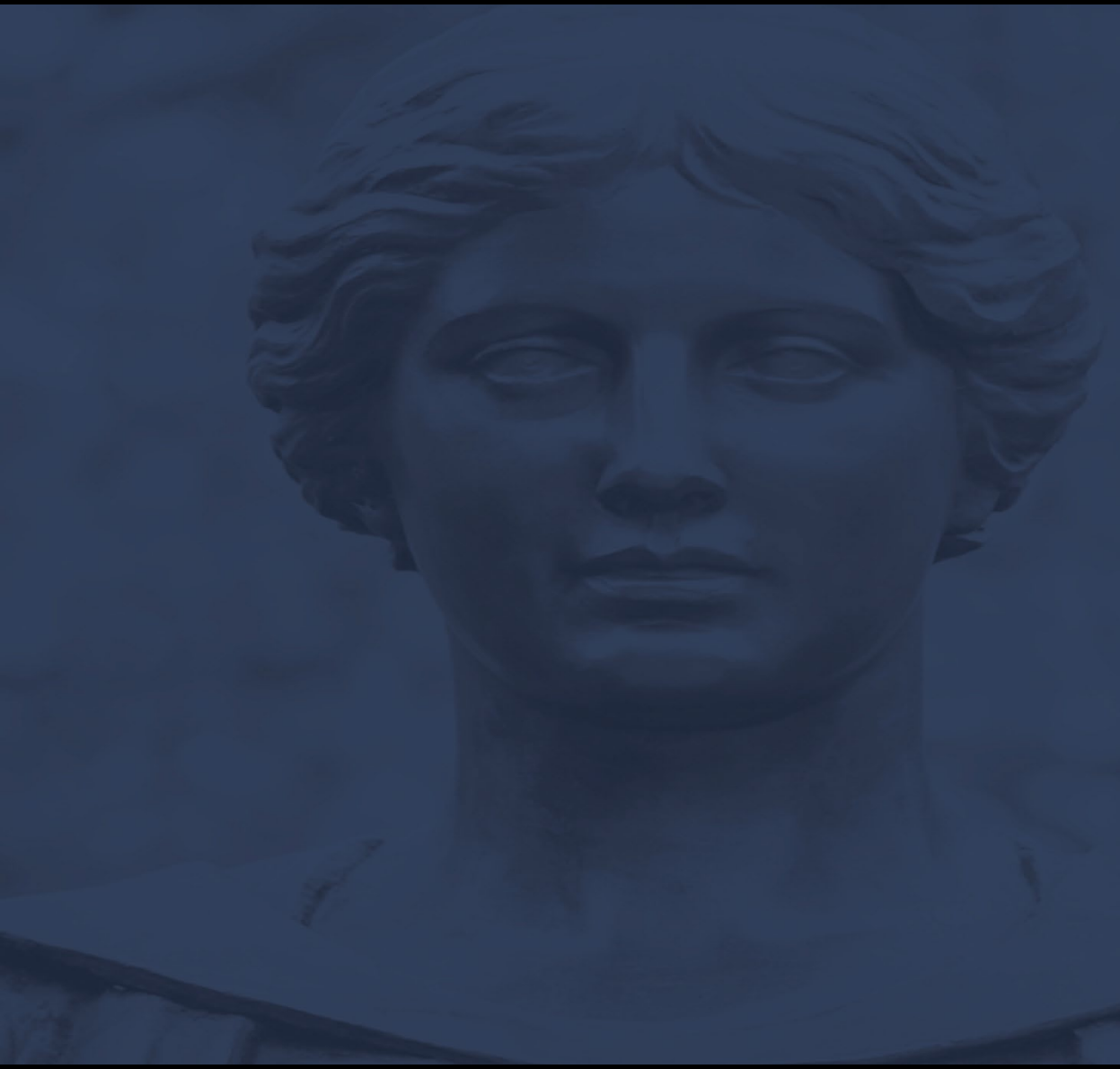
ILLINOIS



Machine Learning in Material Design

Dr. Yumeng Li

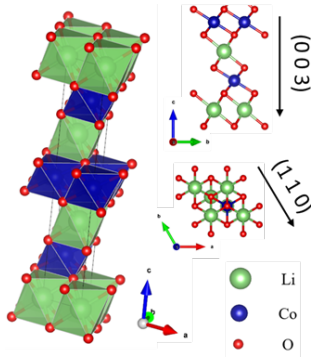
Office: TB 212

Email: Yumengl@illinois.edu

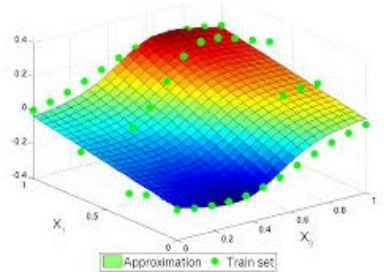
Generic process of building a ML model

Dataset construction

Featurization/ Encoding

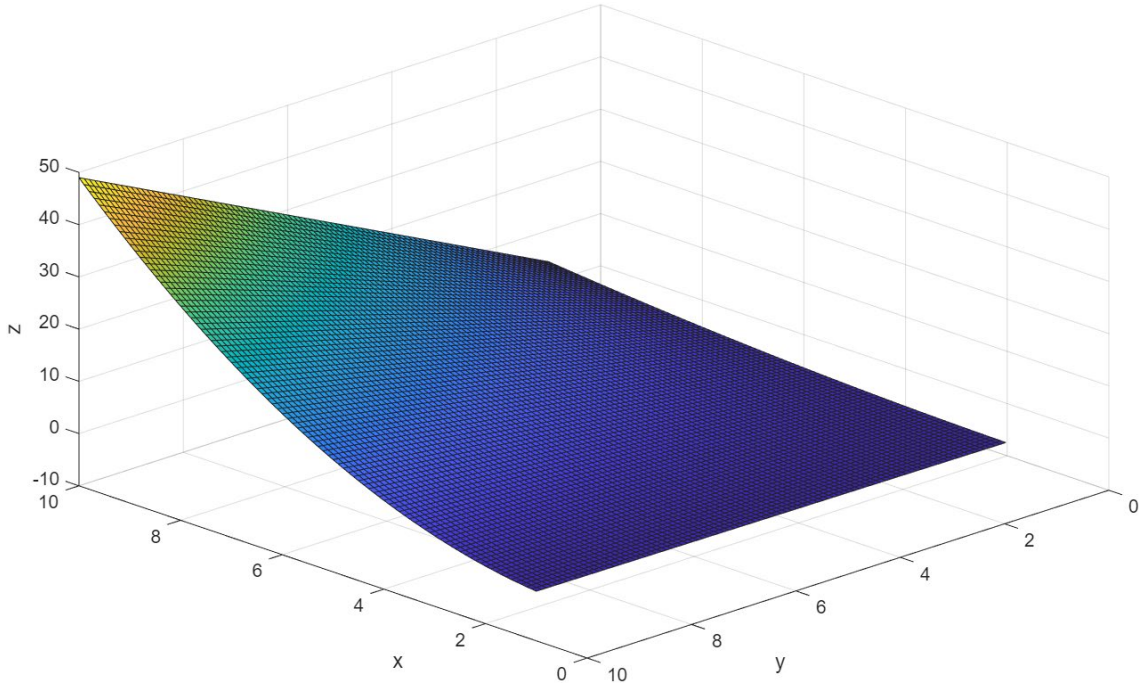


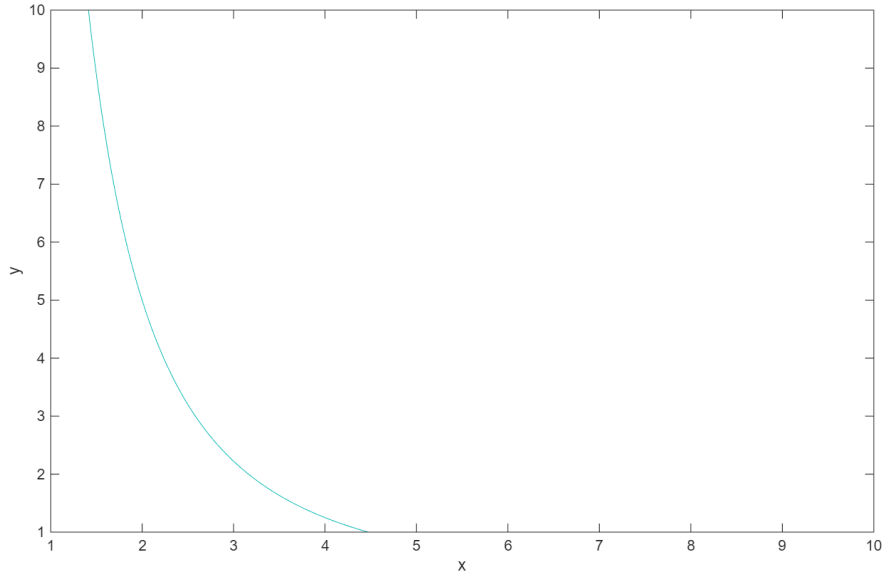
Training/Optimization



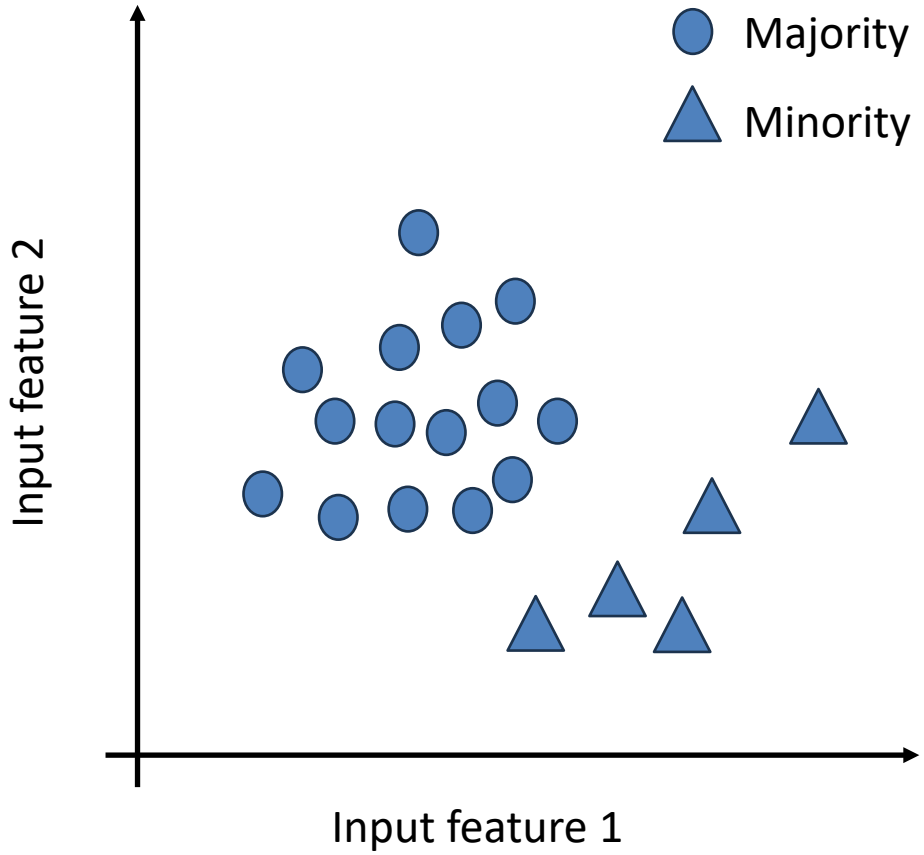
Evaluation

Case study


50
40
30
20
10
0
-10
z
10
8
6
4
2
0
x
0
10
8
6
4
2
0
y


10
9
8
7
6
5
4
3
2
1
y
1
2
3
4
5
6
7
8
9
10
x

Imbalanced data

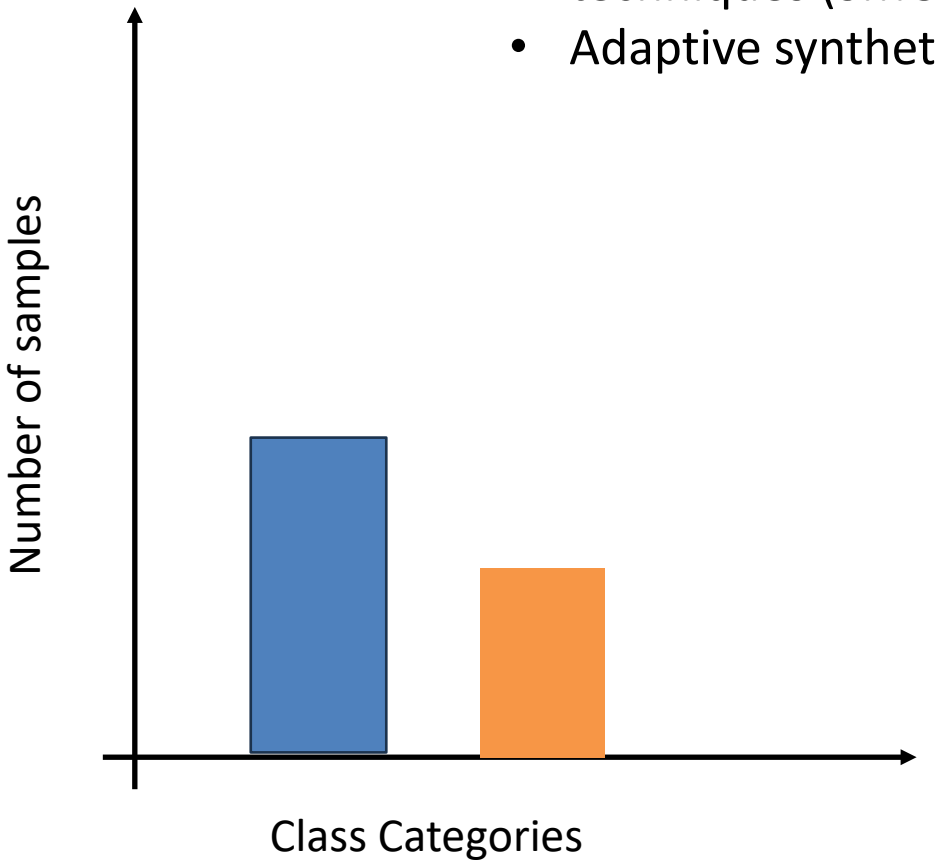


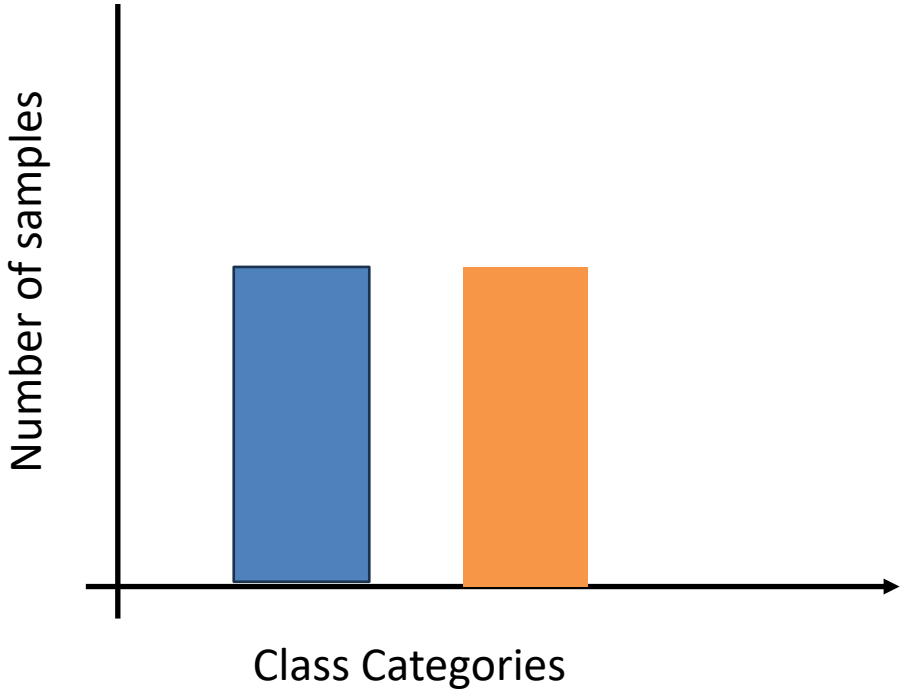
General Approach

- ❑ Data-level approaches
  - Oversampling
  - Undersampling
  - Generative sampling
- ❑ Algorithm-level approaches
  - Cost-sensitive
  - Kernel-based methods
  - Feature reduction
- ❑ Hybrid/ensemble approaches

Oversampling

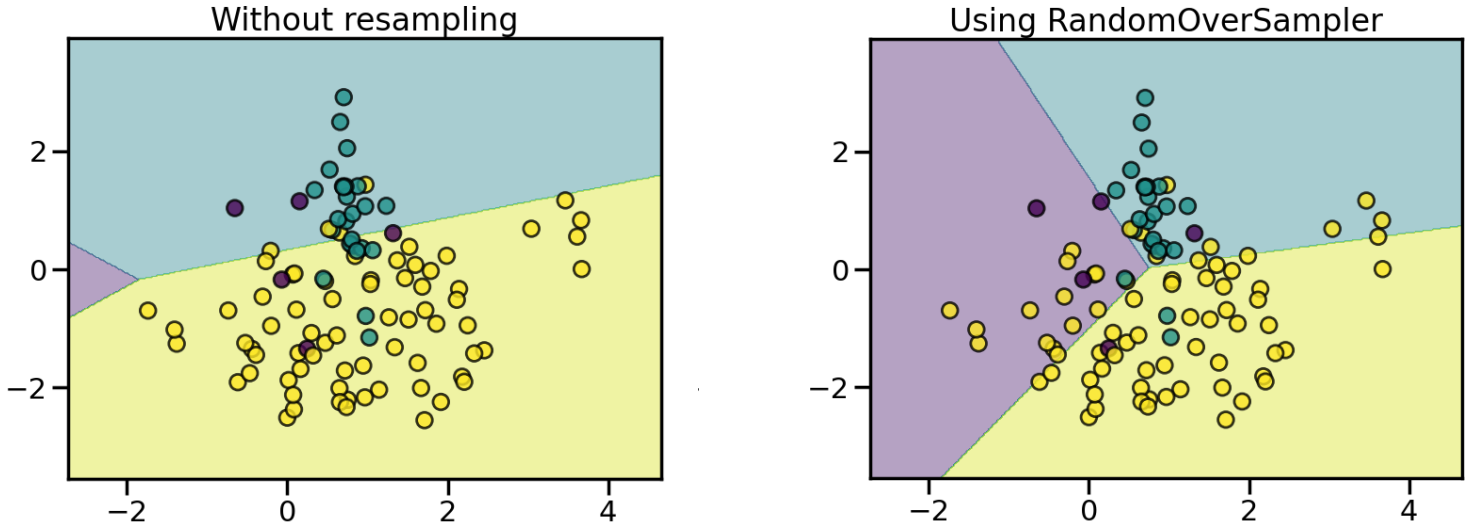
- Oversampling
  - Random oversampling
  - Synthetic minority oversampling techniques (SMOTE) and the variants
  - Adaptive synthetic sampling (ADASYN)



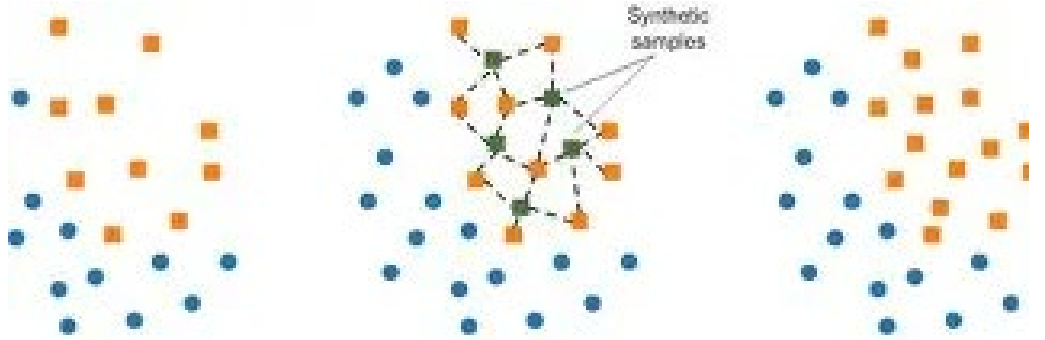


Oversampling

- Random oversampling
  - Simple to implement
  - Can lead to overfit as no new information is introduced and there is lack of the diversity in minority class



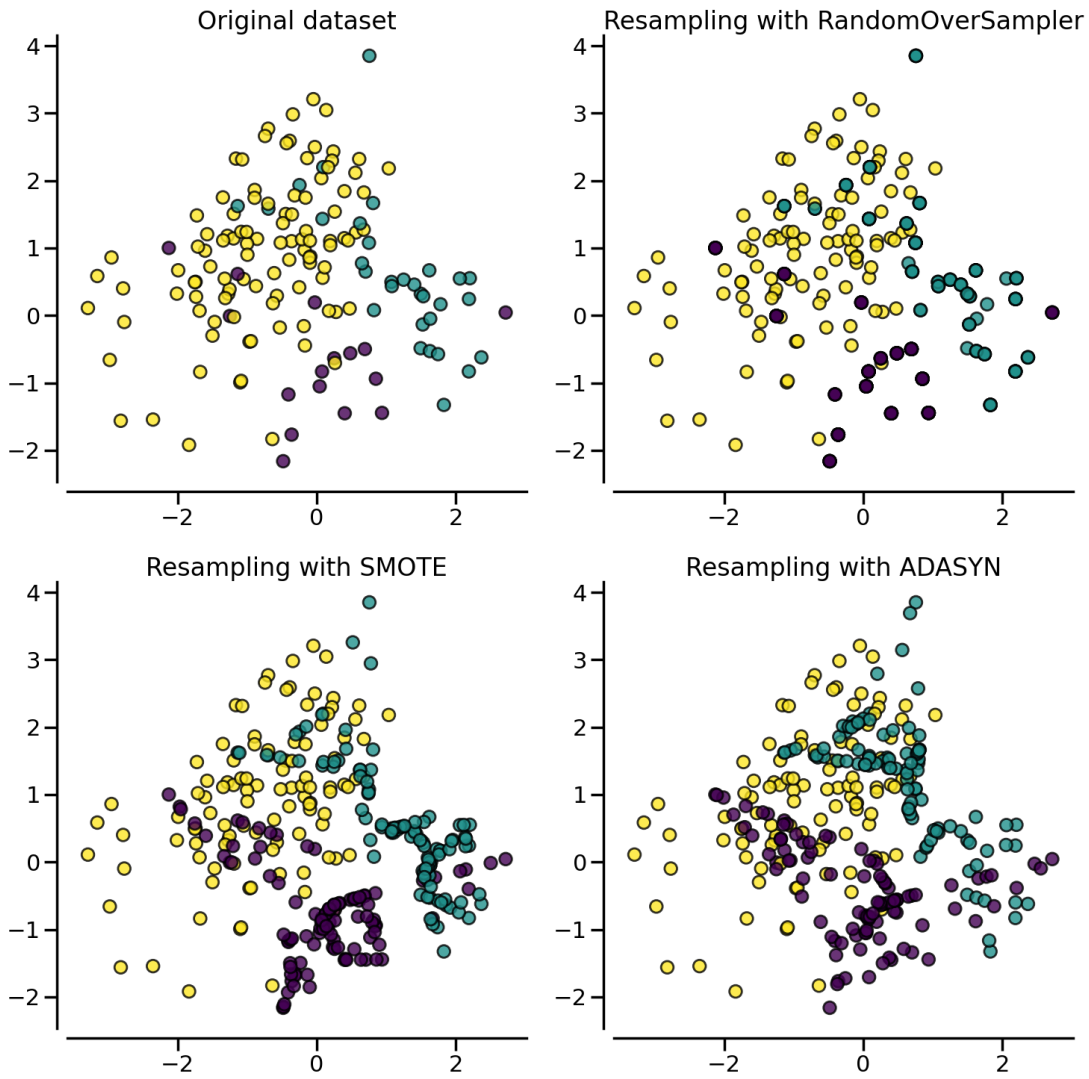
- Synthetic Minority oversampling technique(SMOTE)
  - Generate synthetic minority samples by interpolating between existing minority class instances
  - A lot of variation, Bi-phasic SMOTE, Borline-SMOTE



https://imbalanced-learn.org/stable/over_sampling.html

Oversampling

- Adaptive synthetic sampling (ADASYN)
  - Adaptively generate minority data samples according to their distribution



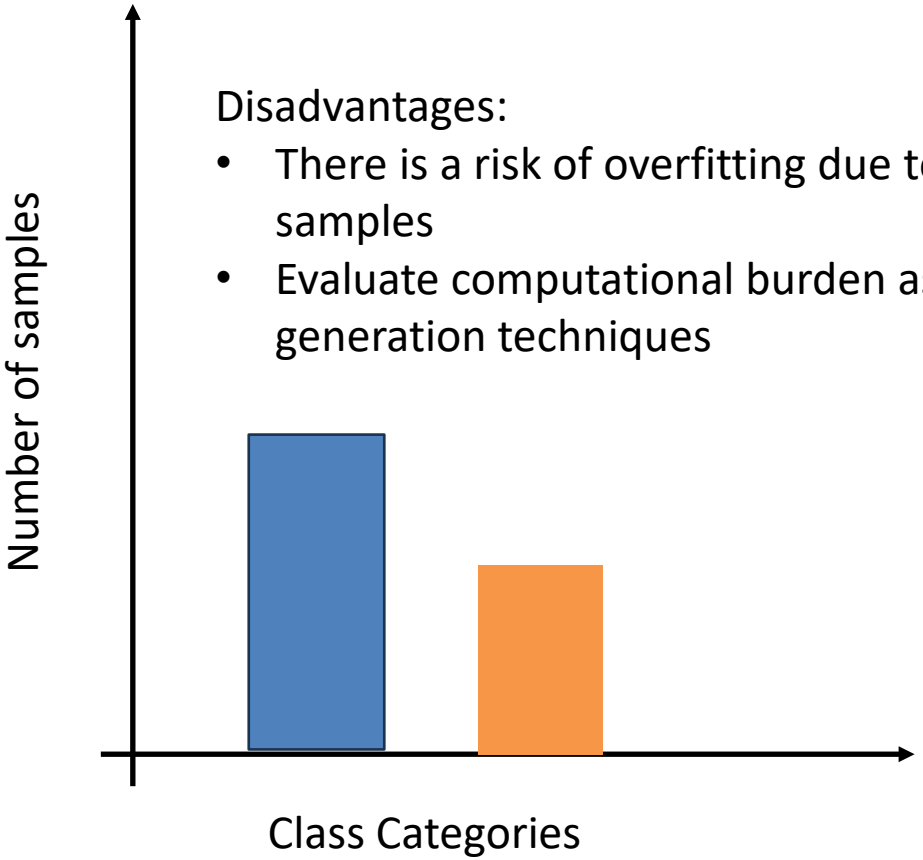
Oversampling

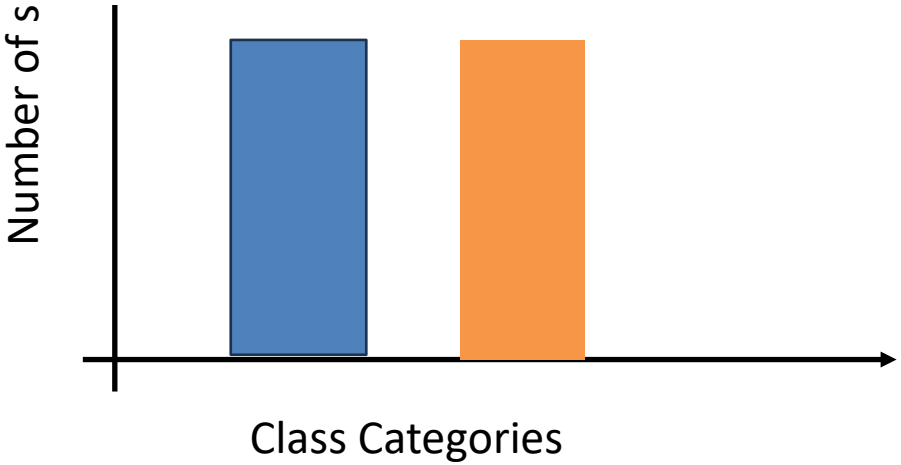
Advantages:

- Increases the model's generalization by enhancing the variability in the minority class
- Safeguards against the loss of essential information in minority class instances

Disadvantages:

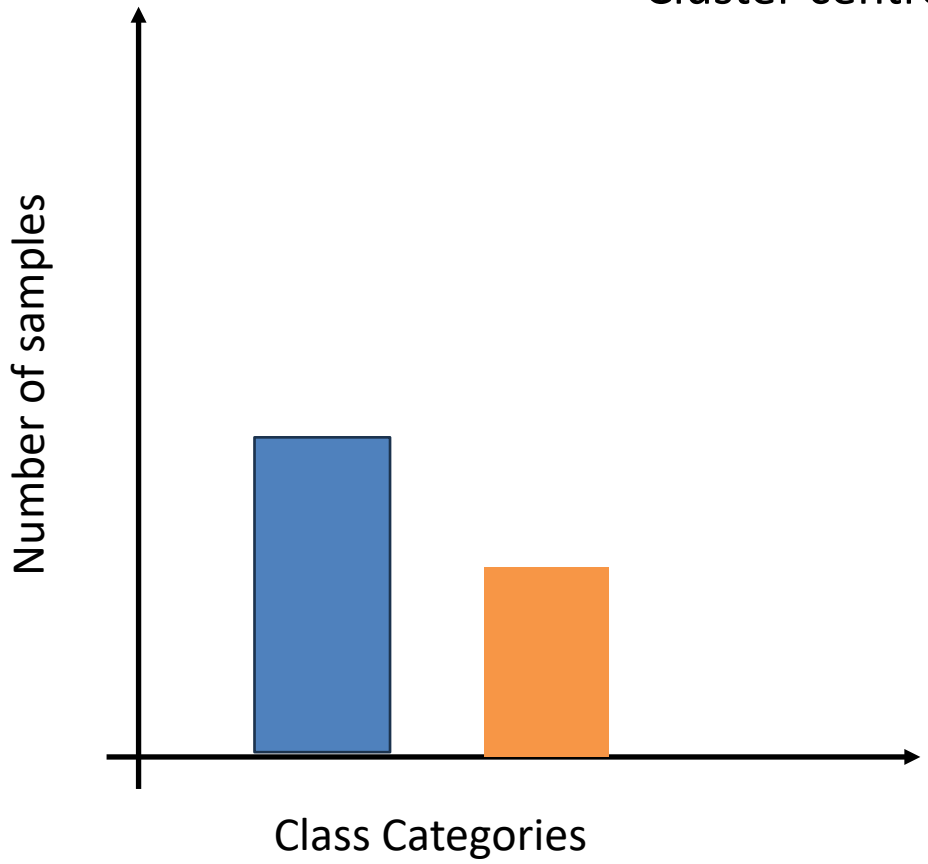
- There is a risk of overfitting due to the noise inherent in the synthetically generated samples
- Evaluate computational burden associated with sophisticated synthetic instance generation techniques

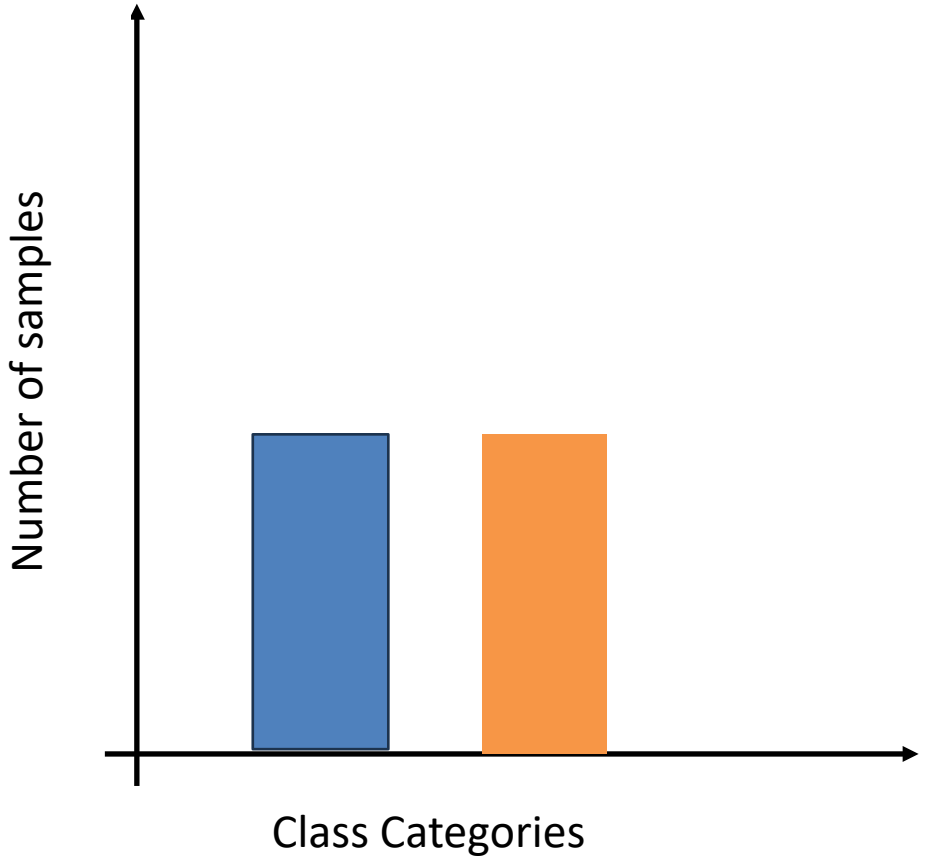




Undersampling

- Undersampling
  - Random undersampling
  - Tomek links
  - Cluster centroids

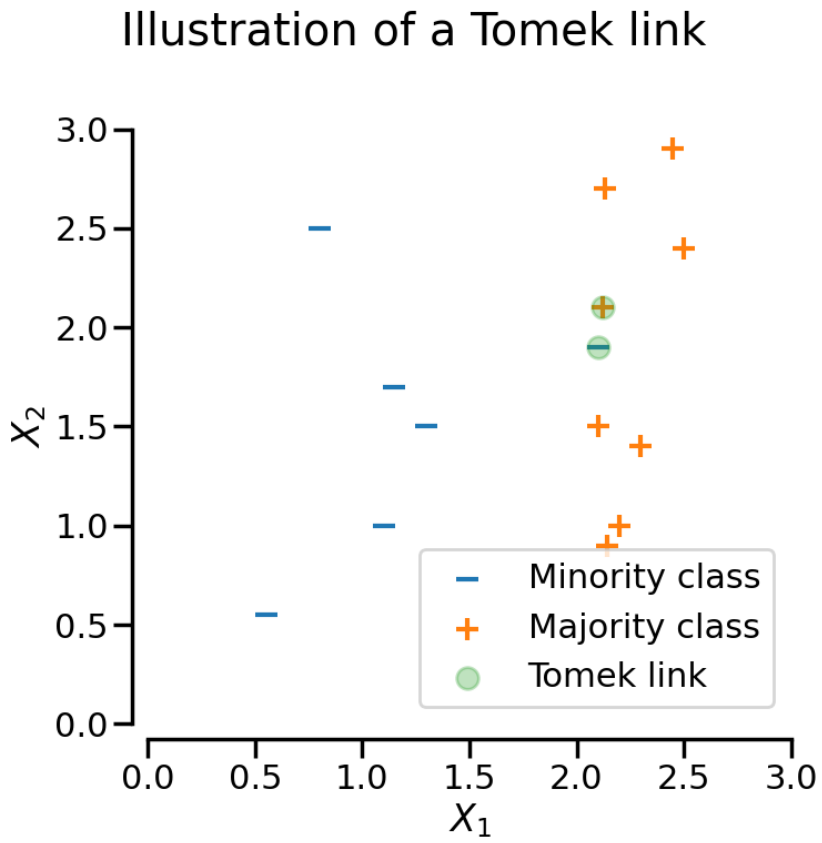




Undersampling

- Random undersampling
  - Simple to implement
  - Accidently remove critical sample from majority classes

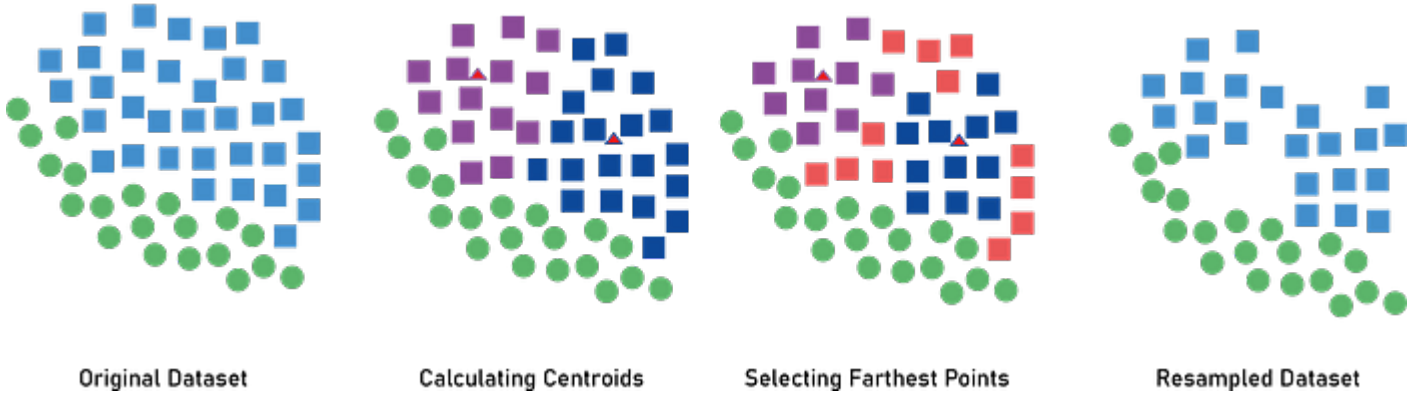
- Tomek links
  - Clean the data by removing the noisy samples or samples that are hard to classify
  - Tomek's link exists when two samples from different classes are the closest neighbors to each other



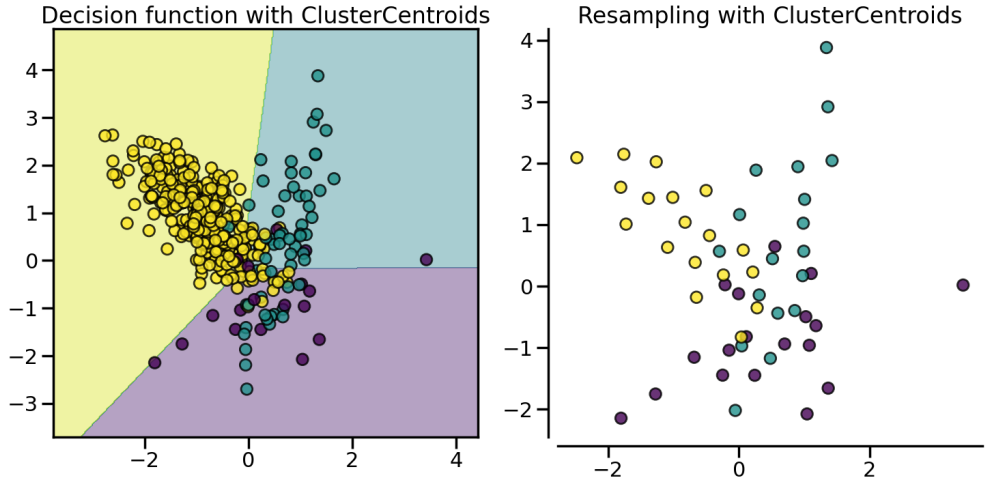
Undersampling

- Cluster centroids
  - Represent the data cluster with a reduced number of samples
  - Need to have the data grouped in to clusters
  - The number of centroids should be set to make sure the representativeness of the original ones

Removing from majority classes



Regenerating the balanced data



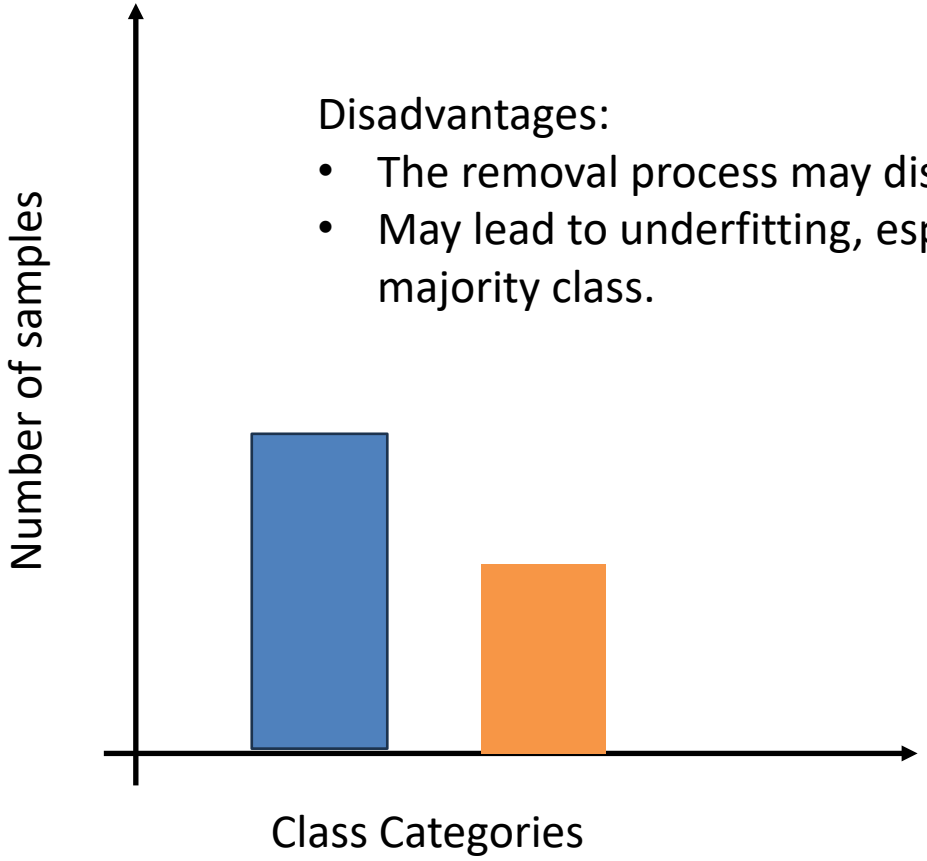
Undersampling

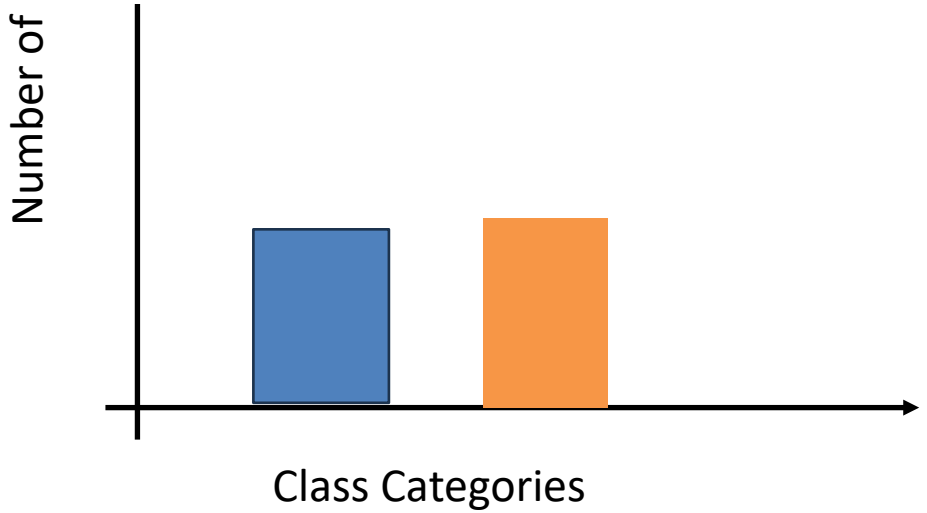
Advantages:

- Reduce the training time and simplify the training process
- Reduce the likelihood of model bias toward the majority class, promoting more fair and balanced learning model

Disadvantages:

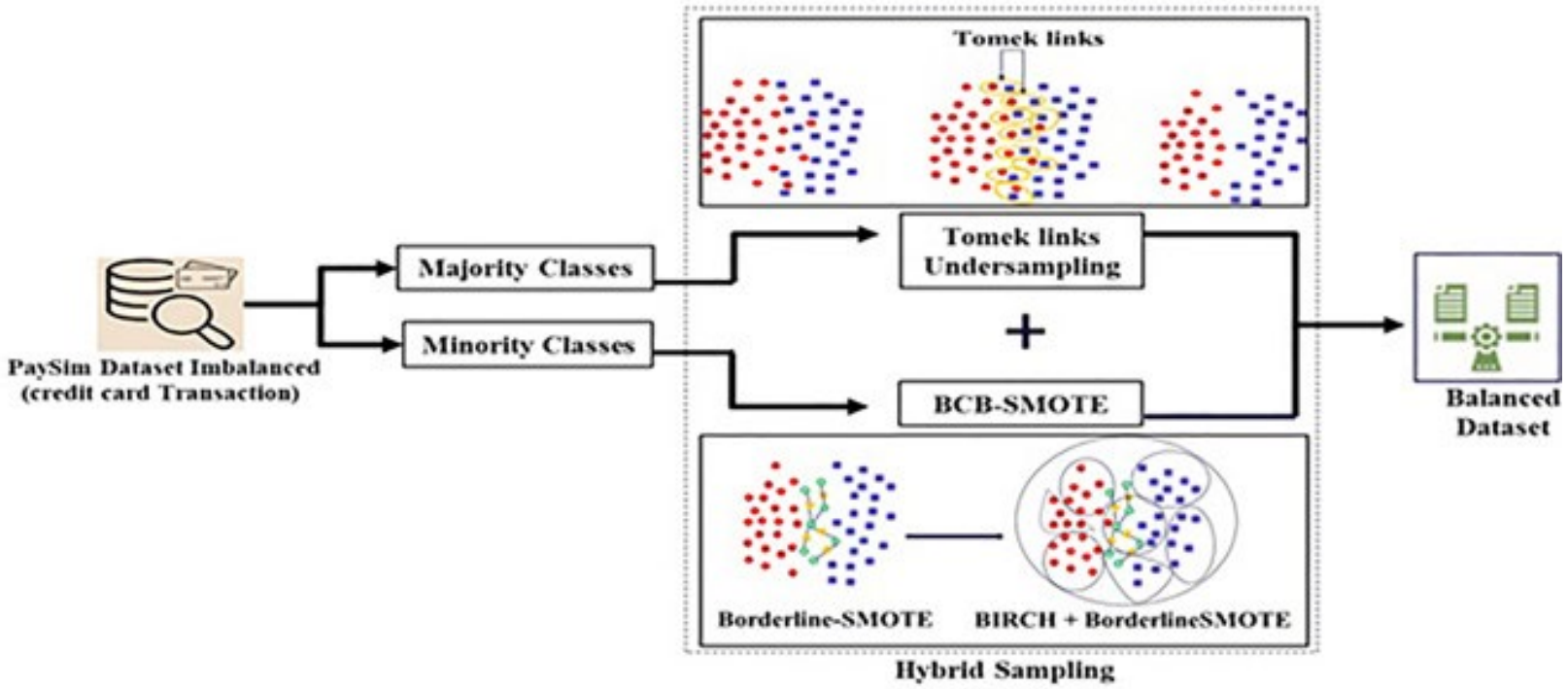
- The removal process may discard crucial instances
- May lead to underfitting, especially if the undersampling reduce the diversity in the majority class.





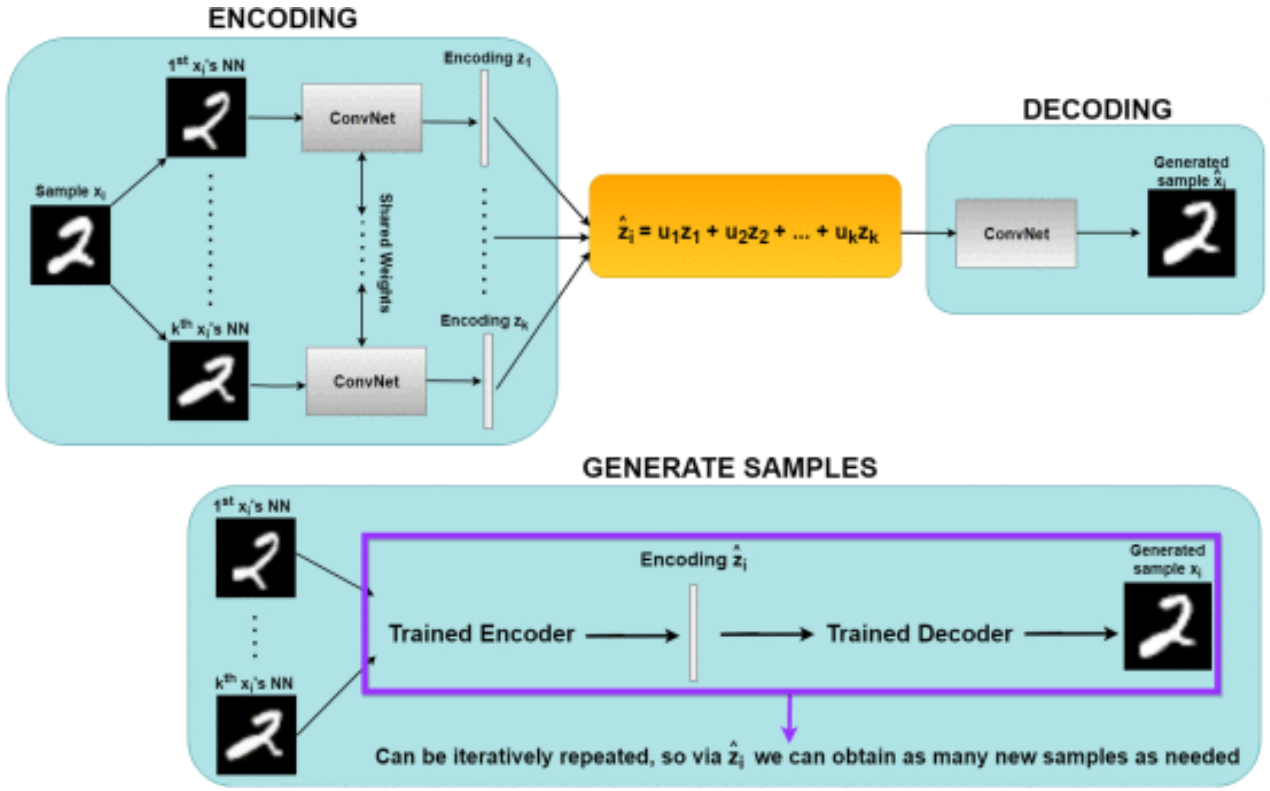
Hybrid sampling method

- Combining resampling technique
  - Combining oversampling and undersampling methods to create a balanced dataset
  - Mitigate the drawbacks of each technique while leveraging their advantages
  - Reducing risk of overfitting
  - Improving model generalization and enhance model performance



M. Alamri and M. Ykhlef, "Hybrid Undersampling and Oversampling for Handling Imbalanced Credit Card Data," in IEEE Access, vol. 12, pp. 14050-14060, 2024,

Generative sampling



E. Troullinou, G. Tsagkatakis, A. Losonczy, P. Poirazi and P. Tsakalides, "A Generative Neighborhood-Based Deep Autoencoder for Robust Imbalanced Classification," in IEEE Transactions on Artificial Intelligence, vol. 5, no. 1, pp. 80-91, Jan. 2024

Algorithm modification

❑ Algorithm-level method

- Class weighting: different weights to different classes. Assign higher weights to the minority class and lower weights to the majority class.
- Cost-sensitive learning: Assign different misclassification costs to classes. Algorithm are trained to minimize the overall cost which can be skewed toward the minority class.
- Threshold adjustment: Control the trade-off between precision and recall in a classification problem.
- Anomaly detection: for extreme class imbalance, the minority class can be treated as anomalies and employ specialized anomaly detection techniques. Examples include One-Class SVM.

❑ Algorithms that can handle imbalanced data

- Below algorithms are considered to be better suited to handling imbalanced data. They partition data into subsets based on class and make independently in each subset.
- Random forest and decision tree
- Support vector machine
- Neural networks

Cost sensitive method

- The goal of cost-sensitive learning is to minimize the cost of a model on a training dataset

- In cost-sensitive learning, a penalty associated with an incorrect prediction is referred as a “cost”.
  - Cost of misclassification or prediction errors
  - Cost of tests or evaluation
  - Cost of teacher or labeling
  - Cost of intervention
  - Cost of unwanted achievements or outcomes from intervening
  - Cost of computation or computational complexity
  - Cost of data collection
  - Cost of human-computer interaction
  - Cost of instability or variance known as concept drift

- Evaluation metrics

Ensemble Method

- Ensemble method is to merge multiple base models to forge a more robust learner for imbalanced data
  - Boosting: train a sequence of learners. Rectify the errors of the predecessors by assigning higher importance to misclassified cases.
  - Bagging/bootstrap aggregating: generate multiple training dataset and train a set of base learners. The predictions of base learners will be aggregated, usually through averaging.
  - Cost-sensitive ensemble:
  - Hybrid

Regression model

- ❑ Assessment metrics
- ❑ Ensemble methods
- ❑ Adaptive sampling